Plant Soil (2006) 288:357–371
DOI 10.1007/s11104-006-9126-z

ORIGINAL PAPER

Two classification methods for developing and interpreting productivity zones using site properties

Nicolás Martín · Germán Bollero · Newell R. Kitchen · Alexandra N. Kravchenko · Ken Sudduth · William J. Wiebold · Don Bullock

Received: 13 January 2006 / Accepted: 6 September 2006 / Published online: 6 October 2006

Abstract Crop performance is often shown as areas of differing grain yield. Many producers utilize simple GIS color ramping techniques to produce visual yield maps with delineated clusters. However, a more quantitative approach such as an unsupervised clustering procedure is generally used by scientists since it is much less arbitrary. Intuitively the yield clusters are due to soil and terrain properties, but there is no clear criterion for the delineation. We compared the effectiveness of two delineation or classification procedures: quadratic discriminant analysis (QDA) and k-nearest neighbor discriminant analysis (k-NN) for the study of how yield temporal patterns relate to site properties. This study used three production fields, one in Monticello, IL, and two in Centralia, MO. Clusters were defined using maize (*Zea mays* L.) and soybean (*Glycine max* (L.) Merr.) yield from three seasons. The k-NN had greater and more consistent successful classification rates than did QDA. Classification success rate varied from 0.465 to 0.790 for QDA while the k-NN classification rate varied from 0.794 to 0.874. This shows that areas of certain temporal yield patterns correspond to areas of specific site properties. Although profiles of site properties differ by crop and production field, areas of consistent low maize yield had greater shallow electrical conductivity ($EC_{shallow}$), than those of consistent high maize yield. Furthermore, areas of consistent high soybean yield had lower soil reflectance than those areas of consistent low yields.

Keywords k-means clustering · Quadratic discriminant analysis · k-nearest neighbor discriminant analysis · Yield temporal patterns · Site properties

N. Martín
Syngenta, Inc., 317 330th Street, Stanton, MN 55018, USA

G. Bollero · D. Bullock (✉)
Crop Sciences, University of Illinois, 1102 South Goodwin Ave., Turner Hall, Urbana, IL 61801-4798, USA
e-mail: dbullock@uiuc.edu

N. R. Kitchen · K. Sudduth
USDA-ARS Cropping Systems and Water Quality Research Unit, University of Missouri, Columbia, MO 65211, USA

A. N. Kravchenko
Department of Crop and Soil Sciences, Michigan State University, East Lansing, MI 48824-1325, USA

W. J. Wiebold
Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA

Introduction

Subdividing large fields into smaller homogeneous areas is appealing to producers. A common assumption is that site properties contribute largely to crop performance and yield and thus one should be able to predict yield with these properties (Fleming et al. 2000; Franzen et al. 2002; Johnson et al. 2003; Chang et al. 2003; Schepers et al. 2004). Note that knowing the yield potential of an area does not always provide sufficient information for management, but there is still interest in the information (Bullock and Bullock 2000). Undeniably the relationships between site characteristics and crop performance in production fields are very complex due to the large effect of, and interactions with, stochastic environmental effects during the production season (Bullock and Bullock 2000).

Yield spatial patterns vary within a field for a given year and from season to season (Jaynes and Colvin 1997). There are areas of consistent yield across space and time and there are areas of non-consistent yield across space and time (Eghball et al. 1999; Taylor et al. 2003). These patterns are further complicated since areas of common yield patterns in the same field can differ by crop. For example, in a maize–soybean rotation an area that consistently produces high maize yields does not necessarily produce high soybean yields (Brock et al. 2005).

The first step in the task of understanding how site properties affect spatial patterns of yield is to develop the yield clusters. Areas of common yield patterns across seasons are well and commonly delineated by a non-hierarchical cluster analysis such as the k-means cluster algorithm, which is an unsupervised classification method with no requirements of previous training areas (Lark and Stafford 1997). This method assumes neither a normal distribution nor homogeneous variances for the input variables (Johnson and Wichern 2002). It iteratively calculates the sum of statistical distances from every observation to the group centroid or multivariate mean. The iterative process finishes when a centroid position is found that minimizes the sum of the distances between each observation and its centroid (Khattree and Naik 1999; Johnson and Wichern 2002).

To classify or predict maize yield clusters using site properties (e.g. soil organic matter (SOM), electrical conductivity (EC), slope, etc.) as explanatory variables, Jaynes et al. (2003) and Ping et al. (2005) proposed the use of discriminant analysis. There are numerous discriminate analysis procedures (Khattree and Naik 1999). The common discrimination procedures assume multivariate normality for the individual populations. We shall refer to those as ordinary discriminant analysis procedures. The ordinary discriminant analysis procedure is based on the estimation of linear or quadratic functions to maximize differences between groups using explanatory variables (Khattree and Naik 1999; Johnson and Wichern 2002). Jaynes et al. (2003) used a quadratic function to predict maize yield clusters using site variables as predictors with variable classification success. Ping et al. (2005) used ordinary discriminant linear analysis to predict areas of high and low average cotton yields using six variables with a classification success of 76.9% for 39 field subsections.

A limitation of ordinary discriminant analysis is that it assumes multivariate normality of the explanatory variables. However, site properties in production fields, such as soil P, soil K or SOM, often have non-normal distributions and heterogeneous variances. If the assumptions of normality or of homogeneity of variances are violated, ordinary discriminant analysis leads to misleading inferences (Khattree and Naik 1999). Data transformation can be attempted, but it is often not sufficient to meet these assumptions. There exist non-parametric discriminant analysis procedures and parametric discriminant analysis procedures which utilize distributions other than multivariate normal, such as kernel density and logistic discriminant analysis, but these are not commonly seen in the applied agriculture literature (Khattree and Naik 1999; Kravchenko et al. 2002).

Another issue to consider is the structure of the data as described by Tominaga (1999). In general, yield data are classified as symmetric if all the members of a yield cluster are contiguous and asymmetric if the members of a cluster type are not contiguous. Tominaga (1999), reports that the ordinary discriminant methods such as QDA perform much better when the data structure is symmetric than when the data are asymmetric. This presents a challenge for grain yield data since the clusters are usually not contiguous. For example, a given field may have four different yield classes (very low, moderately low, moderately high and very high) and thus four types of clusters. But not all of the members of a given yield class are likely to be contiguous. Typically, fields consist of areas of similar grain yield, but the clusters are of differing size, non-contiguous and exhibit substantial embedding, in other words there are clusters within (nested) clusters.

Due to these problems with distributional assumptions and data structure the task of relating yield clusters to field properties is not well suited to ordinary discriminant analysis. We propose that the k-NN discriminant analysis procedure is appropriate for large datasets such as these. The main difference between k-NN discriminant analysis and QDA is that the first method considers the distances between the observation to be classified and its nearest neighbors, while the second method considers statistical distances between the observation to be classified and the centriods (multivariate mean vector) of the classes. This makes k-NN more appropriate for classifying asymmetric data sets which aim to predict the toxicity of new chemical compounds (Tominaga 1999). The k-NN has been applied in other research fields to analyze asymmetric datasets. Malhotra et al. (1999) classified business performance more accurately with k-NN than with QDA. Liu et al. (2003) compared classification success of forest inventories using neural networks and statistical methods. They found that k-NN outperformed the remaining statistical methods. We hypothesize that areas of common maize and soybean yield temporal patterns are related to spatial patterns of site properties and that the yield cluster prediction performance of k-NN discriminant analysis is superior to that of ordinary discriminant analysis.

Materials and methods

The study was carried out in three production fields: Williams North (WN), (Lat. 40.3031, Long. –88.5426), located at Monticello, east central Illinois; GVillo (GV) (Lat. 39.2297, Long. –92.1169) and Field1 (F1) (Lat. 39.2346, Long. –92.1469), both located at Centralia, north-central Missouri. The WN field is 16 ha and its main soils are fine-silty illitic, mesic Typic Endoaquaoll; and fine illitic, mesic Mollic Hapludalf (Officer et al. 2004). The GV field is 14 ha while the F1 field is 18 ha. The main soils in these two fields are fine smectitic, mesic Vertic Epiaqualf; fine-loamy mixed superactive, mesic Typic Hapludalf; fine smectitic, mesic Aeric Vertic Epiaqualf; and fine smectitic, mesic Vertic Albaqualf as dominant soils (Kitchen et al. 1999). The 20-year average precipitation from April to October is 670.3 mm and 667.5 mm in Monticello and Centralia, respectively. During 1997 and 2001 precipitation was less than the 20-year average. Similarly, at Centralia, in 1997 and 1999 precipitation was less than the average. In 1998 precipitation was greater than the 20-year average at both locations. Descriptive statistics of the site characteristics are presented on a field-by-field basis in Table 1. All the fields have been in a maize and soybean rotation for a minimum of 15 years. Crop and field management practices followed the standard management practices for East Central Illinois (Illinois Agronomy Handbook 2003) and Central Missouri (Kitchen et al. 1997).

Data collection

Apparent electrical conductivity surveys were performed using a Veris 3100 sensor cart (Division of Geoprobe Systems, Salina, KS). Since conductivity measures are affected by soil moisture, clay content, soil temperature, and salinity, measurements were done under similar conditions in October of 1999 when the field was at field capacity (Kitchen et al. 2003). The sensor has multiple coulter electrodes in a Wenner arrangement (Kitchen et al. 2003). The electrodes are configured in such way that EC was measured at two depths where 90% of the current drop is observed: from 0 to 30 cm ($EC_{shallow}$), and from 0 to 90 cm (EC_{deep}) (Sudduth et al. 2003). These observations were georeferenced with a differential GPS every 4 m along transects separated by approximately 10 m.

Elevation surveys were carried out in October of 1999 with a Leica 500 RTK DGPS (Leica 500 RTK, Leica, Heerburgg, Switzerland) system at WN and an Astech Z Surveyor RTK (Astech Z, Thales Navegation, Carquefou, France) at GV

Table 1 Site variables descriptive statistics for WN, GV, and F1

Field		Elev. Std[a].	Slope	EC^{b}_{deep}	$EC^{c}_{shallow}$	BR[d]	GR[e]	RR[f]	NIR[g]	SOM[h]	Soil P[i]	Soil K[j]
	Units	m	%	mS m^{-1}	mS m^{-1}					%	kg Mg^{-1}	kg Mg^{-1}
WN	Mean	4.1	2.2	38.6	26.8	0.46	0.55	0.55	0.64	3.80	22.4	135.9
	CV	42.1	37.2	26.8	25.3	5.7	9.2	11.9	11.6	16.4	35.2	13.5
	Skewness	–0.46	–0.10	0.77	1.01	–0.37	–0.28	–0.41	–0.70	1.02	3.90	2.52
	Kurtosis	–0.81	–0.89	0.77	0.66	–0.20	–0.23	–0.22	0.02	0.33	23.10	14.64
GV	Mean	4.1	1.9	25.9	16.6	0.45	0.51	0.48	0.64	2.5	34.3	247.3
	CV	52.1	58.6	41.0	43.9	4.7	6.6	10.0	9.4	9.3	39.6	18.3
	Skewness	–0.03	0.66	0.18	1.32	–0.13	0.01	–0.25	1.57	–0.41	1.56	1.12
	Kurtosis	–1.41	0.44	–0.59	1.46	0.48	0.46	0.92	1.04	–0.67	3.17	0.46
F1	Mean	1.8	0.78	18.4	9.30	0.42	0.49	0.41	0.80	2.4	22.5	193.8
	CV	43.2	35.4	40.0	27.8	4.0	4.3	8.6	9.7	8.9	54.6	15.6
	Skewness*	–0.13	1.24	0.79	1.64	0.35	—0.42	0.17	–0.56	–0.42	2.68	1.53
	Kurtosis*	–0.83	2.73	0.72	5.23	1.05	3.03	1.51	1.90	–0.01	3.01	–1.24

* Skewness and kurtosis absolute values greater than one indicate departures from normality

Abbreviations: [a] standardized elevation, [b] deep electrical conductivity, [c] shallow electrical conductivity, [d] blue soil reflectance, [e] green soil reflectance, [f] red soil reflectance, [g] near infrared soil reflectance, [h] soil organic matter, [i] soil phosphorous, [j] soil potassium

and F1, with 2 cm of vertical and horizontal precision. The GPS equipments were mounted on a four wheel ATV. The observations were arranged as a 10 m semi-regular grid. Elevation was standardized to the lowest position of the field. The resulting grid was used to derive field slope as a percentage (Kravchenko and Bullock 2000).

Soils were sampled in July of 2001 at WN, IL, as a modified grid in 390 locations as described by (Martin et al. 2005). Soils located in GV and F1, were sampled in a grid pattern in October of 1995 as described by Officer et al. (2004) in 208 locations for GV, and 466 locations for F1. State and commercial laboratories determined soil organic matter (SOM) (Carlo-Erba analyzer, Carlo-Erba, Italy), (Bray-P1) and K(Mehlich N°3) (Officer et al. 2004).

Bare soil reflectance was acquired in late April and early May of 2001 and 2002 using a multispectral sensor, mounted on a satellite following a sun synchronous orbit at 681 km of altitude (IKONOS, Space Imaging Thornton, CO). The sensor divided the spectrum into four bands: blue (BR), with center in 480.3 nm and a range of 71.3 nm; green (GR), with center in the 550.7 nm and a range of 88.6 nm; red (RR) with center of 664.8 nm and a range of 65.8 nm; and near infrared (NIR), with center in 805 nm and a range of 94.5 nm. The image provider performed image geometric and spectral calibrations.

Table 2 Cluster number, R^2, for yield clusters produced by using yields of multiple seasons, for corn and soybean at WN, GV and F1

Field	Crop	Season	Clusters	R^2
WN	Corn	1996, 1998, 2000	4	0.58
	Soybean	1997, 1999, 2001	4	0.67
GV	Corn	1997, 2000, 2002	3	0.58
	Soybean	1998, 1999, 2001	3	0.47
F1	Corn	1997, 1999, 2001	4	0.66
	Soybean	1998, 2000, 2002	4	0.55

Soybean and maize yields were obtained using combines equipped with yield monitors (Ag Leader Technology, Ames, IA) in six consecutive seasons, three for maize and three for soybean, as seen in Table 2. Grain data was recorded on 1-second intervals and corrected to 13 and 15.5% moisture, in October and September, for soybean and maize, respectively. Unreliable yield observations including positioning errors, abrupt changes in combine speed, and grain flow were removed following the guidelines of Kitchen et al. (2003). Yields were transformed into standard deviation units.

All the variables were interpolated to the same spatial scale as the IKONOS images (4 ×4 m) using regionalized variable theory (ESRI 2001) according to Goovaerts (1997). Isotropic semivariograms were estimated for all the variables using spatial lags of between 8 and 15 m. The semivariogram models we selected using cross validation error mean and standard error as well as plots comparing estimated versus predicted values. The obtained spatial structures were applied in ordinary kriging interpolations to produce multivariate datasets arranged in regular grids of 4 ×4 m.

Data analysis

Descriptive statistics of the site characteristics, obtained with the UNIVARIATE procedure of SAS (SAS Inst. 2002). For each dataset observation of the 4 ×4 m grid (x_1, x_2,... x_n), crop yield data from three season years defined an (X_1, X_2,...X_n) vector. The k-means cluster analysis was performed with FASTCLUS procedure of SAS (SAS Inst. 2002). Observations were grouped according to their Euclidian distances to the group centroid.

$$d_g^2(x_1,x_2) = (X_1 - X_2)'(X_1 - X_2), \qquad (1)$$

where d^2 is square distance between observation x_1 and x_2, (Johnson and Wichern 2002; Khattree and Naik 1999). This distance measure is appropriate for datasets of unequal variances and non-zero covariances.

The clustering algorithm creates clusters by minimizing the distances between observations from the same cluster and by maximizing the distances between observations of different clusters (Khattree and Naik 1999). The k-mean clustering procedure defines as many clusters as requested . The optimum number of groups was selected based on the coefficient of determination (R^2), and the cubic clustering criterion (Khattree and Naik 1999). As cluster number increases, the R^2 increases logarithmically while the cubic clustering criterion decreases exponentially. When the R^2 and the cubic clustering criterion values stabilize, the number of clusters is optimum (Khattree and Naik 1999).

The resulting clusters define areas of distinctive crop performance for three seasons for each crop. Yields clusters in areas with the highest yield performance across the seasons were defined as cluster 1. Then clusters of decreasing yield performance were labeled cluster 2, cluster 3 and cluster 4.

The agreement between maize and soybean grain yield clusters was measured by the Kappa index of agreement (KIA) (Stafford et al. 1996; Ping et al. 2005) using the FREC procedure of SAS (SAS Inst. 2002). The KIA varies between 0 (no association) and 1 (perfect association). This index is based on contingency tables that compare how many observations in a given soybean grain yield cluster correspond to a similar maize grain yield cluster

$$\mathrm{KIA} = \frac{P_\mathrm{o} - P_\mathrm{e}}{1 - P_\mathrm{e}}, \qquad (2)$$

where P_o = cumulative proportion of actual matches for each cluster over the total number of observations, P_e = cumulative proportion of expected matches for each cluster over the total number of observations.

The occurrence of a certain yield cluster was predicted using the site variables (SOM, K, P, soil reflectances, $\mathrm{EC_{deep}}$, $\mathrm{EC_{shallow}}$, elevation and slope), for each location where a yield cluster was defined using two alternative methods of discriminant analysis: quadratic ordinary discriminant analysis (QDA) and k-NN discriminant analyses (k-NN).

Quadratic discriminant functions predict location group membership probability to a certain group as

$$\hat{P}(c|X_i) = \frac{q_c|S_c|^{-\frac{1}{2}}.\exp(-\frac{1}{2}D_{ic}^2)}{\sum_{g'=1}^{k} q_{c'}|S_{c'}|^{-\frac{1}{2}}.\exp(-\frac{1}{2}D_{ic'}^2)}, \qquad (3)$$

where D_{ic}^2 is the Mahalanobis distance, which considers the variance covariance matrix, between each observation and a c or c' group centroid or multivariate mean (Johnson and Wichern 2002). As the distance between a point i and the centroid c decreases with respect to the other centroid c', the probability of belonging to c increases (Huberty 1994; Johnson and Wichern 2002). These distances are proportional to the output of quadratic classification function made for each group Q_{ic}

$$-2Q_{ic} = \ln S_c^{-1} + D_{ic}^2 - 2\ln q_c, \qquad (4)$$

Thus, as result of the quadratic classification function Q_{ic}, the distance to the group centroid D_{ic}^2 decreases and $\hat{P}(c|X_i)$ increases. The Q_{ic} for each observation is estimated as,

$$Q_{ic} = c + b_c X_i + X_i' A_c X_i, \tag{5}$$

where c is the intercept, **b** is the vector of linear coefficients and **A** is the matrix of quadratic coefficients for a group g (Khattree and Naik 1999; Johnson and Wichern 2002).

The k-NN operates based on multivariate distances between observations that are measured using classification variables (slope, elevation, $\mathrm{EC_{deep}}$, $\mathrm{EC_{shallow}}$, BR, GR, RR, NIR). Location membership to a given group (cluster 1, cluster 2, etc.) will depend upon the yield cluster membership of the four nearest neighbors in terms of the multivariate Mahalanobis distance. These multivariate distances depend on the site properties for each location. Membership probability for a location i in a certain group (or yield cluster) c, is indicated as (Huberty 1994),

$$\hat{P}(c|X_i) = \frac{m_c}{\sum_{c'=1}^{k} m_{c'}}, \tag{6}$$

where X_i is the initial population of i locations, m_c is the number of units belonging to a group c, and k is the number neighbors used to predict group membership (Huberty 1994). The observation will belong to the group where the highest probability is observed.

Site characteristics variances were shown to be heterogeneous across groups, thus linear and canonical discriminant analysis were unsuitable. Therefore, quadratic discriminant analysis was selected. Normal score transformations were applied to guarantee the normal distribution of the classification variables as suggested by Goovaerts (1997) and Kravchenko et al. (2002).

The prediction success of k-NN was compared to that of the QDA (Jaynes et al. 2003). The successful observation classification into different clusters of each discrimination method was evaluated with the FREC cross-validation procedure of SAS (SAS Inst. 2002). This method consists of setting aside one observation to test the classification function obtained with the remaining observations (Huberty and Lowman 1997). This process is repeated for all the observations in the dataset.

Prediction performance was evaluated comparing the number of successful predictions n_c by each method with the expected success if the observations were assigned by random chance

$$e_c = t_c q_c, \tag{7}$$

where t_c is the number of observations in group g and q_c is the probability of belonging to group c. The increase in prediction performance, I, between k-NN and QDA and random chance was estimated as

$$I = \frac{Hn_c - He_c}{1 - He_c} 100, \tag{8}$$

where Hn_c and He_c correspond to the proportion of successful classification and expected random chance assignment, respectively (Huberty 1994). This index measures how much the classification success improves when random observation assignment to a given cluster is replaced by any discriminant analysis method.

Results and discussion

Yield cluster identification

For all year and crop combinations, the k-means clustering procedure successfully separated areas of consistent crop yields across seasons (Table 2). According to the coefficient of determination (R^2) clusters explained from 0.47 to 0.67 of the yield variation across years.

Field WN and F1 had four maize and four soybean clusters while GV had three maize and three soybean clusters (Table 2). When crop yield patterns of the individual clusters were investigated across seasons, cluster 1 had the highest consistent yield performance with the exception of maize at GV in 2000 (Fig. 1). Clusters 2 and 3 had intermediate yield performance for WN and F1. Finally, cluster 4 at WN and F1 and cluster 3 at GV consistently had the lowest yield. For all fields and crops, the yield of the lower quartile of cluster 1 was greater than the yield of the upper

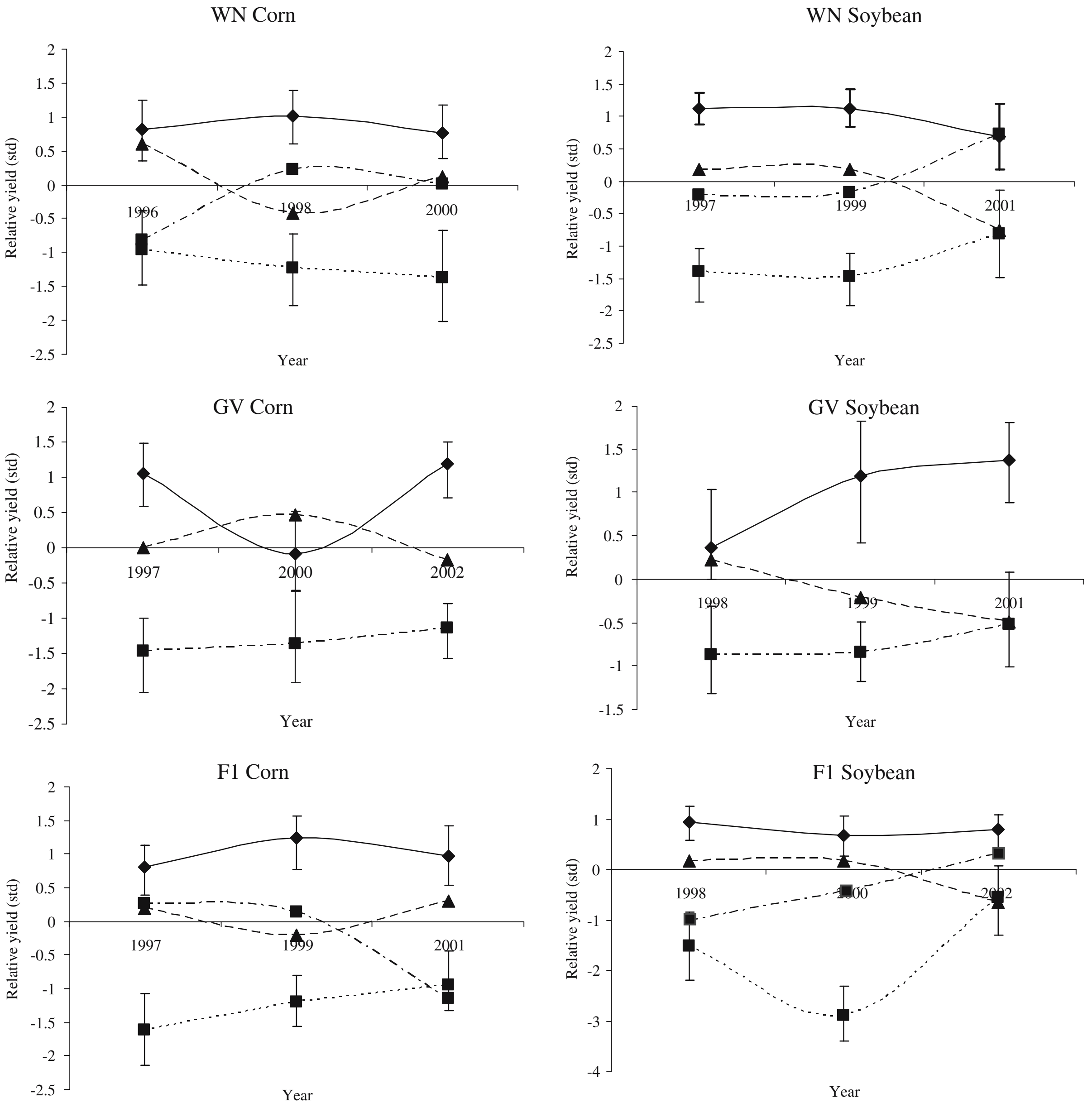

Fig. 1 Cluster yields in standard deviations for corn (**a**, **c**, **e**) and soybean (**b**, **d**, **f**), at WN (**a** and **b**), at GV (**c** and **d**) and at F1 (**e** and **f**). Standardized yields were represented as—◆— for cluster 1, – –▲– – for cluster 2, – · ■ · – for cluster 3, and ·······■······· for cluster 4. The bars, indicates the quartile boundaries for cluster 1 and cluster 4

quartile of cluster 4 at WN and F1 and cluster 3 at GV (Fig. 1). Note, that clusters 1 and 4 at WN and F1 and cluster 1 and 3 at GV are consistent high and low yielding, respectively and thus stable, while clusters 2 and 3 at WN and F1 and cluster 2 at GV are unstable and exhibit a range of yield over time. This outcome agrees with the categories presented by Blackmore et al. (2003). They proposed that within a given production field there are stable-high-yielding areas, stable-low-yielding areas, and unstable areas. Brock et al. (2005) described their clusters according to their average yield across seasons in high, medium-high, medium-low, and low yield.

Alternatively, Jaynes et al. (2003) associated yield cluster not with average yield but with different landscape positions.

Spatial agreement between maize and soybean clusters

There were different levels of agreement between clusters of yield patterns for maize and soybean but overall KIA values were lower than 0.4. In WN, the proportion of matching clusters (P_o) was 0.386, and KIA was 0.213. In GV the agreement of clusters was the best with a P_o of 0.575, and a KIA of 0.328. Conversely, in F1 maize and soybean cluster 1 had a P_o of 0.344, and KIA of 0.132. Similarly, Brock et al. (2005) using weighed KIA found diverse levels of agreement between soybean and maize yield clusters that ranged from 0.06 to 0.34.

The results suggest that clusters of maize grain yield do not always correspond to the same field areas of clusters soybean grain yield, and areas of stable low maize yield do not always correspond to areas of stable low soybean yield. Brock et al. (2005) reported stronger associations between wheat (*Triticum aestivum* L.) yield clusters versus corn and soybean clusters than between corn and soybean clusters only. They suggested that there were differences in the spatial structure of the limiting factors for corn and soybean. Sadras and Calviño (2001) reported that under water supply restrictions maize yield decreased more than soybean yield. In addition, Kaspar et al. (2004) indicated that in eroded areas of lower water and nutrient supply maize yield decreased more than soybean yield. Conversely, they also found that under excessive water supply soybean yield decreased more than maize yield.

Table 3 Classification performance of quadratic and k-NN discriminant analysis at WN, GV, F1 production fields for corn and soybean in a cluster-by-cluster basis: number of observations (t_c), expected number of successfully classified observations by random chance (e_c), proportion successfully classified observations by random chance (He_c), successfully classified observations (n_c), proportion of successfully classified observations (Hn_c), improvement in classification success compared with random cluster assignment (I_c)

Field	Crop	Cluster	t_c (obs.)	e_c (obs.)	He_c (prop.)	QDA			k-NN		
						n_c (obs.)	Hn_c (prop.)	I_c (%)	n_c (obs.)	Hn_c (prop.)	I_c (%)
WN	Corn	Cluster 1	2576	711	0.076	1969	0.764	74.5	2186	0.849	83.7
		Cluster 2	2583	715	0.077	1393	0.539	50.1	2100	0.813	79.7
		Cluster 3	2494	668	0.072	870	0.349	29.9	1955	0.784	76.7
		Cluster 4	1666	298	0.032	886	0.532	51.7	1349	0.81	80.4
	Soybean	Cluster 1	2322	578	0.062	749	0.323	27.8	1884	0.811	79.8
		Cluster 2	2556	700	0.075	1150	0.45	40.5	2029	0.794	77.7
		Cluster 3	2618	736	0.079	1726	0.659	63.0	2044	0.781	76.2
		Cluster 4	1823	357	0.038	777	0.426	40.3	1441	0.790	78.2
GV	Corn	Cluster 1	1941	489	0.064	1471	0.758	74.2	1708	0.88	87.2
		Cluster 2	4352	2463	0.320	3861	0.887	83.4	4037	0.928	89.4
		Cluster 3	1400	255	0.033	1013	0.724	71.5	1110	0.793	78.6
	Soybean	Cluster 1	1298	219	0.029	750	0.578	56.6	1026	0.79	78.4
		Cluster 2	3489	1584	0.206	2603	0.746	68.0	3049	0.874	84.1
		Cluster 3	2906	1098	0.143	2094	0.721	67.5	2600	0.895	87.8
F1	Corn	Cluster 1	3684	866	0.055	2675	0.726	71.0	3314	0.900	89.4
		Cluster 2	6520	2719	0.174	4935	0.757	70.6	5925	0.909	89.0
		Cluster 3	2419	375	0.024	841	0.348	33.2	1881	0.778	77.3
		Cluster 4	3023	583	0.037	2423	0.802	79.4	2749	0.909	90.5
	Soybean	Cluster 1	4365	1218	0.078	2535	0.581	54.6	3857	0.884	87.4
		Cluster 2	6604	2787	0.178	5474	0.829	79.2	5802	0.879	85.3
		Cluster 3	3897	970	0.062	1586	0.407	36.8	3209	0.823	81.1
		Cluster 4	780	39	0.002	427	0.547	54.6	589	0.755	75.4

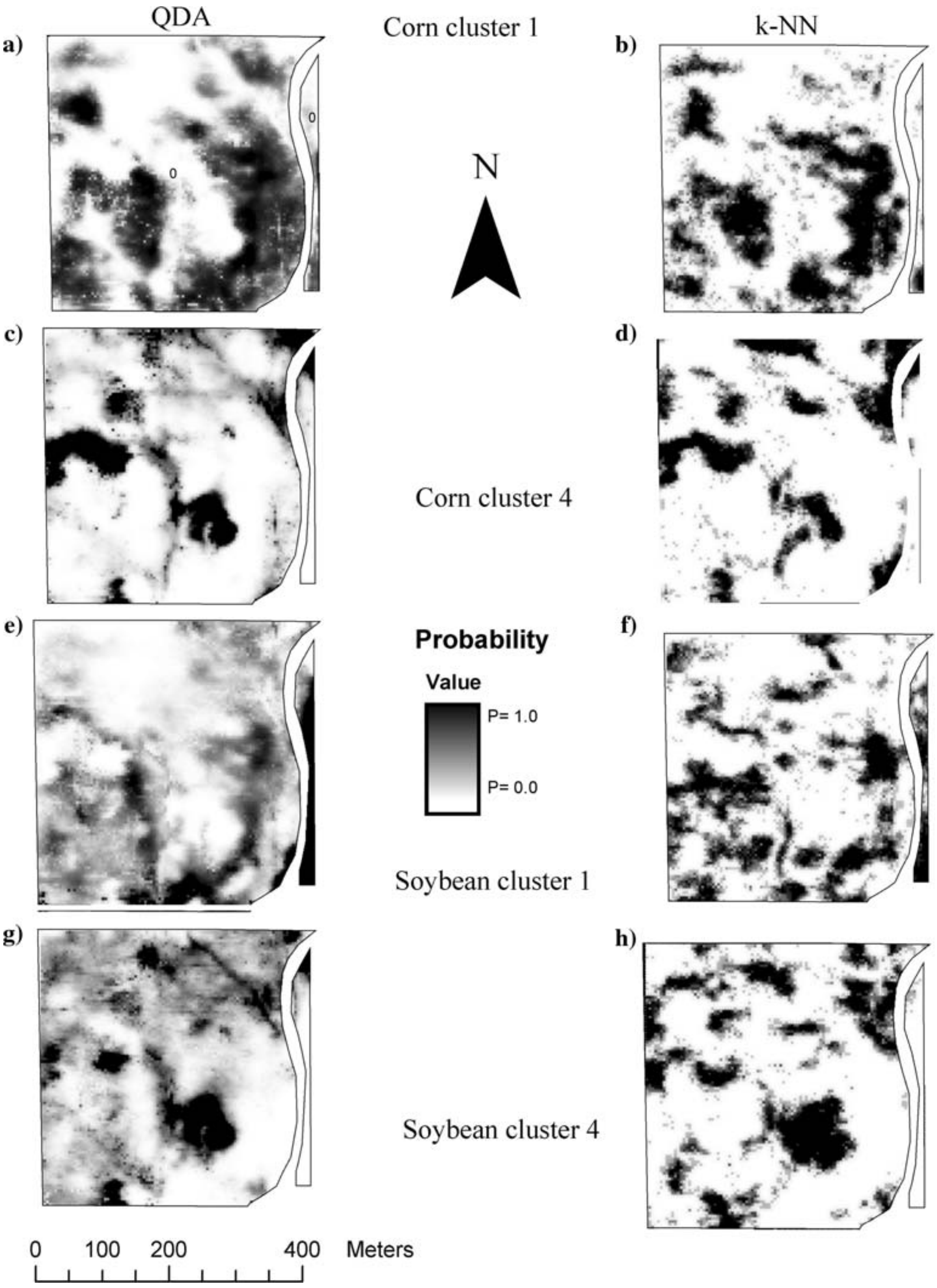

Fig. 2 Maps of yield cluster membership probability at WN for corn (**a**), (**b**), (**c**) and (**d**) and soybean (**e**), (**f**), (**g**), (**h**). Cluster 1 predicted with QDA (**a**) and (**e**); and with k-NN (**b**) and (**f**). Cluster 4 membership probability predicted with QDA (**c**) and (**g**); n with k-NN (**d**) and (**h**)

Yield cluster prediction

For both prediction methods (i.e. QDA and k-NN), site characteristics had higher probabilities of classifying grain yield clusters than that of random chance classification, as measured by the *I* index (Table 3). These results imply that areas of common yield pattern relate to site characteristics.

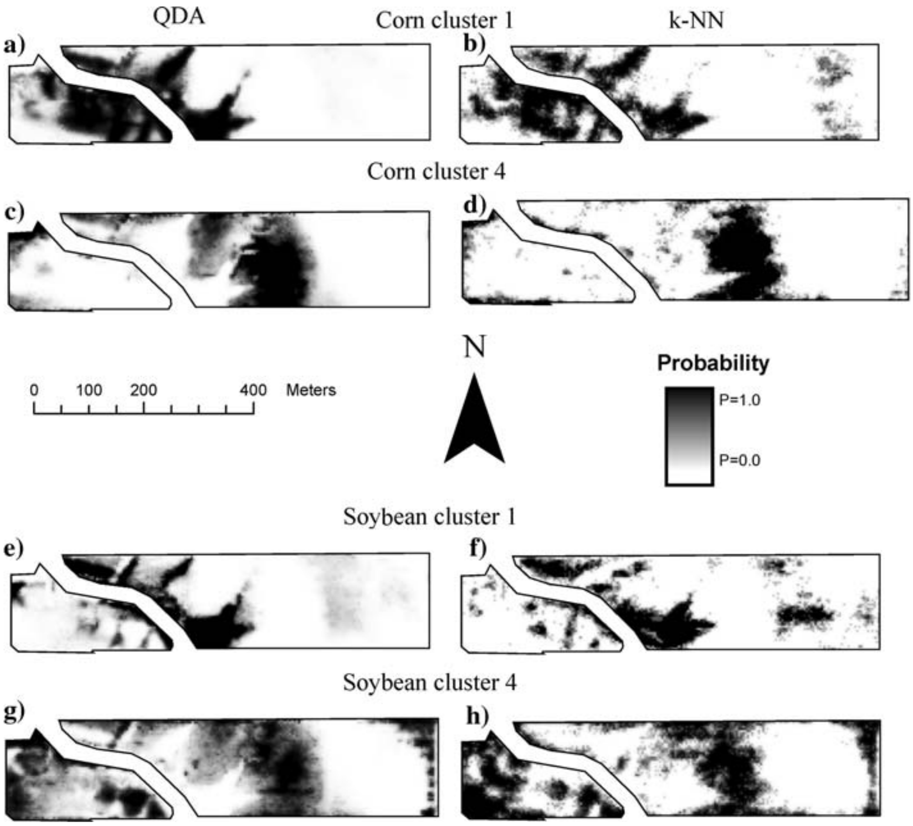

Fig. 3 Maps of yield cluster membership probability at GV for corn (**a**), (**b**), (**c**) and (**d**) and soybean (**e**), (**f**), (**g**), (**h**). Cluster 1 predicted with QDA (**a**) and (**e**); and with k-NN (**b**) and (**f**). Cluster 4 membership probability predicted with QDA (**c**) and (**g**); n with k-NN (**d**) and (**h**)

This outcome agrees with the results presented by Jaynes et al. (2003), Brock et al. (2005); and Ping et al. (2005), who found the site properties were related to yield clusters. Both methods also estimate the probability of a certain location to belong to a certain yield cluster using site variables. These probabilities were mapped in Figs. 2, 3, and 4.

The proportion of successful classified observations (*Hn*) and the improvement in classification success rates compared to random assignment (*I*) were greater for k-NN than QDA (Table 3). When classification performance was evaluated on a by cluster basis, the k-NN presented a more consistent classification performance than that of QDA (Table 3). When QDA was used, the classification success (Hn_c) for the 22 clusters had a mean of 0.612 and CV of 28.2%. The classification success varied from 0.323 (*I* = 27.8%) for soybean cluster 1 in WN to 0.887 (*I* = 83.4%) for maize cluster 2 (Table 3). These results are consistent with the uncertain classification results obtained by Jaynes et al. (2003). They reported classification success rates that ranged form 0.142 to 1. They improved poor classification success by merging the clusters with the lowest success rate.

Overall clusters, the classification success of QDA had a mean of 0.612 and a CV of 28.21%. The classification success for k-NN had a mean of 0.832 and CV of 6.36%. The classification success varied from 0.323 (I_c = 27.8%) to 0.887 (I_c = 83.4%). The classification success varied from 0.755 (I_c = 75.4%) in cluster 4 for soybean at WN to 0.928 (I_c = 89.4%) in cluster 2 for maize at GV (Table 3). The increase in classification success when k-NN is compared with QDA was an average 24.3% and it was statistically

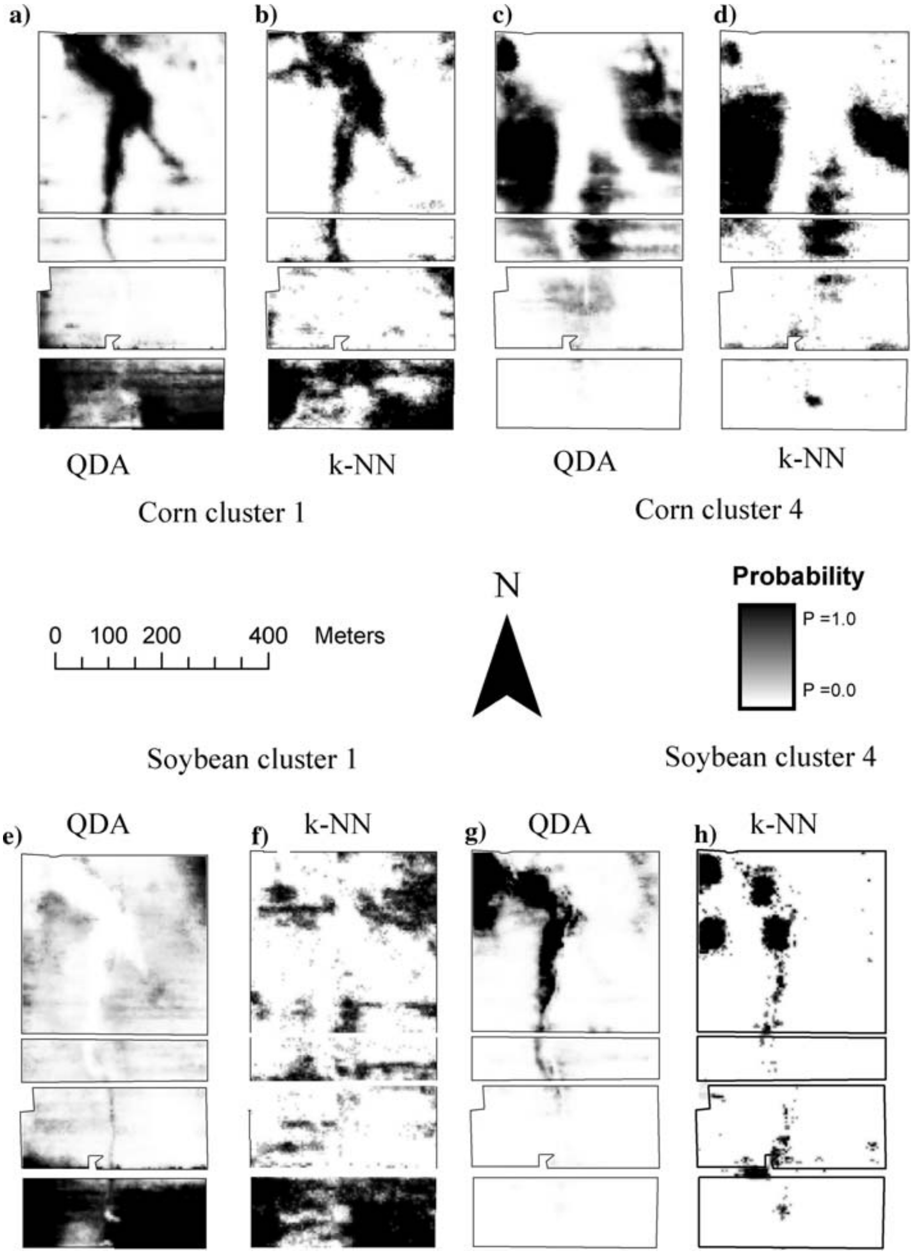

Fig. 4 Maps of yield cluster membership probability at F1 for corn (**a**), (**b**), (**c**) and (**d**) and soybean (**e**), (**f**), (**g**), (**h**): Cluster 1 predicted with QDA (**a**) and (**e**); and with k-NN (**b**) and (**f**). Cluster 4 membership probability predicted with QDA (**c**) and (**g**); and with k-NN (**d**) and (**h**)

significant ($t = 7.945$). This clearly shows the superiority of in classification performance of k-NN over QDA discriminant analysis. The differences in classification success were consistent for different cluster types.

Site characteristics associated with yield clusters

Site characteristic profiles in Figs. 5 and 6 show the site attribute mean and quartiles for

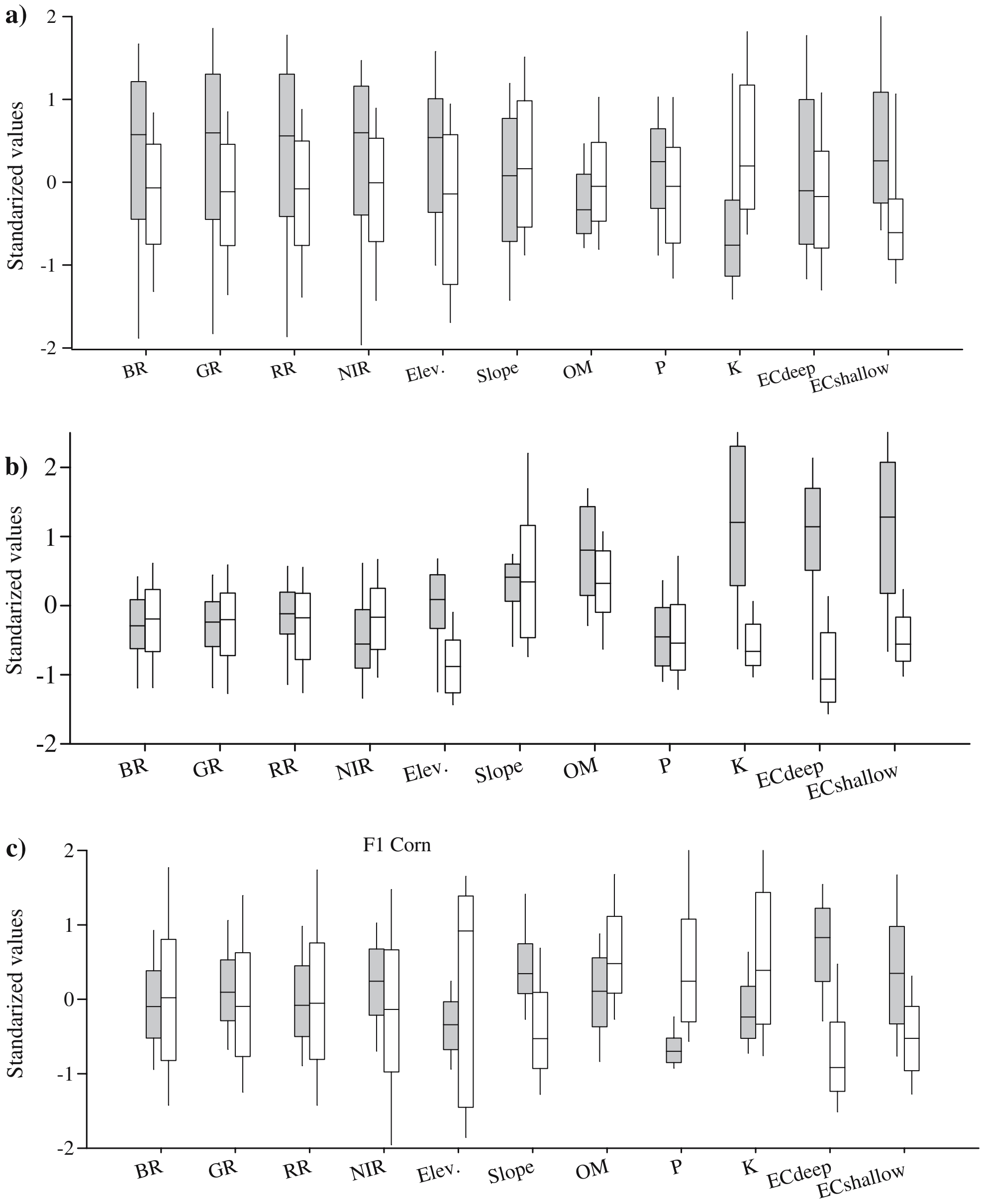

Fig. 5 Site variables boxplots for corn yield cluster 1 (□) and cluster 4 (▨) or cluster 3 for WN (**a**), GV (**b**), and F1 (**c**) production fields

observations successfully classified by k-NN discriminant into highest and lowest yield clusters. These profiles show how site characteristics relate to specific yield clusters (Figs. 5, 6) showing the actual distribution of site attributes of successfully classified observations.

For both crops in WN, areas of stable low yields (cluster 4) had greater soil reflectance in all the spectral bands associated with low SOM (Figs. 5a, 6a). This response agrees with results from multiple studies (Fernandez et al. 1988; Chen et al. 2000; Fox et al. 2003). Soil SOM influences soil reflectance in a wide spectrum region (from 400 to 2500 nm) (Ben-Dor 2002). In uphill areas, erosion removed the topsoil and this reduced SOM and increased soil reflectance. The

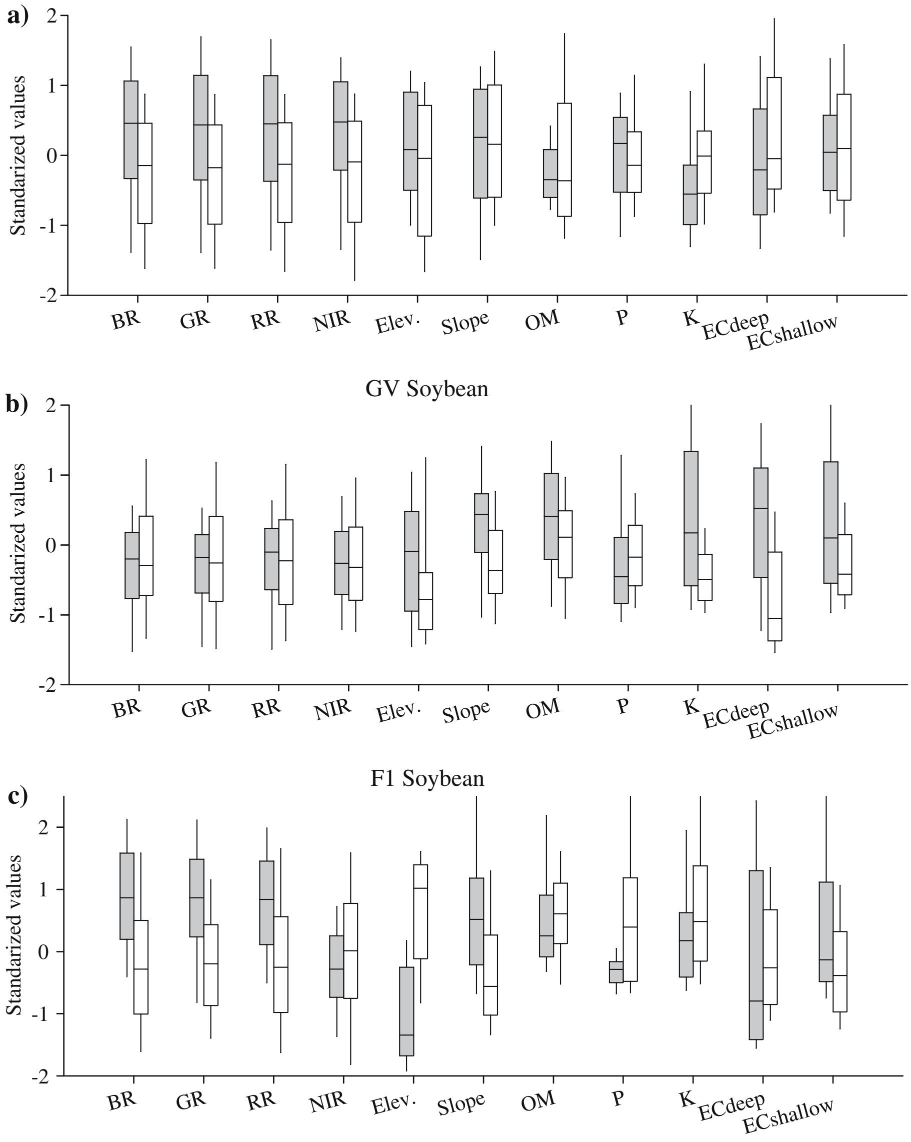

Fig. 6 Site variables boxplots for soybean yield cluster 1 (□) and cluster 4 (▨) or cluster 3 for WN (**a**), GV (**b**), and F1 (**c**) production fields

lower SOM, eroded areas reduces soil water and nutrient supply capacity, reducing crop yield both crops. On the contrary, areas of consistent higher yields are located in the lower positions of the field, higher SOM and K. These areas are consistent with the areas of high SOM levels subject to rill erosion deposition close to the drainage ways as described by Officer et al. (2004). Maize yield clusters 4 and 1 differ in $EC_{shallow}$ while soybean yield clusters 4 and 1 present similar values and dispersion. Areas of higher $EC_{shallow}$ are associated with areas of higher clay content in the surface (Sudduth et al. 2003) and may have restrictions in plant-available water supply that affected crops differentially.

The higher levels of agreement between maize and soybean yield clusters in GV can be explained with site characteristics profile (Figs. 5b, 6b). For both crops in GV, clusters 4 showed higher EC_{deep} and $EC_{shallow}$ values than clusters 1. The EC_{deep}, $EC_{shallow}$ and exchangeable K indicate the presence of shallow claypans (Kitchen et al. 2003; Officer et al. 2004). Claypans close to the surface impede root development and make crops prone to water stress and lower nutrient supply (Kitchen et al. 1999).

For both crops in F1, cluster 4 areas had higher slopes than cluster 1 areas (Figs. 5c, 6c). This field gave the greatest disagreement between maize and soybean yield cluster spatial patterns. On one hand, maize yield cluster 4 had larger EC_{deep} and $EC_{shallow}$ than that of maize yield cluster 1. As in GV, these areas correspond to shallow claypans layers (Kitchen et al. 2003). On the other hand, soybean yield cluster 4 areas had larger BR, GR, RR, and slope as well as lower SOM and elevation than soybean cluster 1. As observed in WN, the increase in soil reflectance relates to SOM losses in eroded areas associated with lower plant-available soil water and nutrient supply, and thus lower soybean yield.

These results indicate that each cluster had a specific profile of site variables according to the crop and the field. Nevertheless, $EC_{shallow}$ observations differ in maize yield cluster 4 and cluster 1 in all the fields. Maize yield were consistently lower in areas with shallow claypan depth (Kitchen et al. 2003) or areas were erosion removed the topsoil layer (Officer et al. 2004). In WN and F1 variables BR, GR, RR, and NIR were consistently related to areas of lower soil SOM and lower soybean yields. In GV there is a clear agreement in the areas of high yield maize and high soybean yield. This can be attributed to the stronger influence of a shallow claypan layer in this field, which is related to high $EC_{shallow}$ and EC_{deep} observations in cluster 4 areas.

Conclusions

This study concludes that site characteristics relate to stable, low and high yield across seasons. When the relationships were evaluated, the k-NN classified yield clusters more accurately than did QDA, which had a poorer and inconsistent prediction performance for some yield clusters.

We speculate that complex soil plant relationships cannot always be described by a symmetric data structure where yield clusters can be separated by a single discriminant function, such as in GV where both discriminant analyses preformed well. Thus, in other fields it may be necessary to use a different discrimination algorithm appropriate for non-symmetric data structures such as k-NN to improve classification success. k-NN, depends on the observation neighborhood and thus it can address complex data structures (Tominaga 1999). In other words, areas of consistent yield have a symmetric structure in some fields and non-symmetric in others (WN). It is interesting to notice that the clustering procedure and the prediction with k-NN are based on multivariate statistical distances. Thus this analysis demonstrates that areas of similar profiles of site characteristics present similar yield patterns across seasons.

Although this study indicates that field-by-field and crop-by-crop analyses can provide a better insight on how site characteristics relate to yield performance, areas of consistent greater maize yield are located in areas of low $EC_{shallow}$. Also, areas of consistent greater soybean yields were located in areas lower soil reflectance, in two of the three fields studied. Moreover, relative to the similar soil spectral relationship, future studies should consider the use of panchromatic images, which have single spectral band and greater spatial resolutions. Given these results, future studies must address, if yield clusters are appropriate to fit site-specific management practices

References

Ben-Dor E (2002) Quantitative remote sensing of soil properties. Adv Agron 75:173–243

Blackmore S, Godwin RJ, Fountas S (2003) The analysis of spatial and temporal trends in yield map data over six years. Biosys Eng 84:455–466

Brock A, Brouder SM, Blumhoff G, Hofmann BS (2005) Defining yield-based management zones for corn–soybean rotations. Agron J 97:1115–1128

Bullock DS, Bullock DG (2000) Economic optimality of input application rates in precision farming. Prec Agric 2:71–101

Chen F, Kissell DE, West LT, Adkins W (2000) Field-scale mapping of surface soil organic carbon using remotely sensed imagery. Soil Sci Soc Am J 64:746–753

Chang JY , Clay DE, Carlson CG, Clay SA, Malo DD, Berg R, Kleinjan J, Wiebold W (2003) Different techniques to identify management zones impact nitrogen and phosphorus sampling variability. Agron J 95:1550–1559

Eghball B, Hergert GW, Lesoing GW, Ferguson RB (1999) Fractal analysis of spatial and temporal variability. Geoderma 88:349–362

ESRI (2001) Using ArcGIS geostatistical analyst. ESRI, Redlands,CA

Fernandez RN, Schulze DG, Coffin DL, Van Scoyoc GE (1988) Color, organic matter, and pesticide adsorption relationships in a soil landscape. Soil Sci Soc Am J 52:1023–1026

Fleming KL, Westfall DG, Wiens DW, Brodahl MC (2000) Evaluating farmer defined management zone maps for variable rate fertilizer application. Precision Agric 2:201–215

Fox GA , Sabbagh GJ, Searcy SW (2003) Radiometric normalization of multi-temporal images based on image soil lines. Trans ASAE 46:851–859

Franzen DW, Hopkins DH, Sweeney MD, Ulmer MK, Halvorson AD (2002) Evaluation of soil survey scale for zone development of site-specific nitrogen management. Agron J 94:381–389

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, NY

Huberty CJ (1994) Applied discriminant analysis. John Willey Publications, NY

Huberty CJ, Lowman LL (1997) Discriminant analysis via statistical packages. Educ Psychol Meas 57:759–784

Illinois Agronomy Handbook (23rd edn) (2003) University of Illinois extention, Urbana-Champaign, IL

Jaynes DB, Colvin TS (1997) Spatiotemporal variability of corn and soybean yield. Agron J 89:30–37

Jaynes DB, Kaspar TC, Colvin TS, James DE (2003) Cluster analysis of spatiotemporal corn yield patterns in an Iowa field. Agron J 95:574–586

Johnson RA, Wichern DW (2002) Applied multivariate analysis, 3rd edn. Prentice Hall, NJ

Johnson CK, Mortensen DA, Wienhold BJ, Shanahan JF, Doran JW (2003) Site-specific management zones based on soil electrical conductivity in a semiarid cropping system. Agron J 95:303–315

Kaspar TC, Pulido DJ, Fenton TE, Colvin TS, Karlen DL, Jaynes DB, Meek DW (2004) Relationship of corn and soybean yield to soil and terrain properties. Agron J 96:700–709

Khattree R, Naik DN (1999) Applied multivariate statistics with SAS software. SAS Institute, Cary, NC; Wiley, NY

Kitchen NR, Blanchard PE, Hughes DF, Lerch RN (1997).Impact of historical and current farming systems on groundwater nitrate in northern Missouri. J Soil Water Conserv 52(4):272–277

Kitchen NR, Sudduth KA, Drummond ST (1999) Soil electrical conductivity as a crop productivity measure for claypan soils. J Prod Agric 12:607–617

Kitchen NR, Drummond ST, Lund ED, Sudduth KA, Buchleiter GW (2003) Soil electrical conductivity and topography related to yield for three contrasting soil-crop systems. Agron J 95:483–495

Kravchenko AN, Bollero GA, Omonode RA, Bullock DG (2002) Quantitative mapping of soil drainage classes using topographical data and soil electrical conductivity. Soil Sci Soc Am J 66:235–243

Kravchenko AN, Bullock DG (2000) Correlation of corn and soybean grain yield with topography and soil properties. Agron J 92:75–83

Lark RM, Stafford JV (1997) Classification as a first step in the interpretation of temporal and spatial variation of crop yield. Ann Appl Biol 130:111–121

Liu CM, Zhang LJ, Davis CJ, Solomon DS, Brann TB, Caldwell LE (2003) Comparison of neural networks and statistical methods in classification of ecological habitats using FIA data. For Sci 49:619–631

Malhotra MK, Sharma S, Nair SS (1999) Decision making using multiple models. Eur J Oper Res 114:1–14

Martin NF, Bollero GA, Bullock DG (2005) Associations between field characteristics and soybean plant performance using canonical correlation analysis. Plant Soil 273:39–55

Officer SJ, Kravchenko A, Bollero GA, Sudduth KA, Kitchen NR, Wiebold WJ, Palm HL, Bullock DG (2004) Relationships between soil bulk electrical conductivity and the principal component analysis of topography and soil fertility values. Plant Soil 258:269–280

Ping JL, Green CJ, Bronson KF, Zartman RE, Doermann A (2005) Delineating potential management zones for cotton based on yields and soil properties. Soil Sci 170:371–385

Sadras VO, Calviño PA (2001) Quantification of grain yield response to soil depth in soybean, maize, sunflower, and wheat. Agron J 93:577–583

SAS Institute (2002) SAS user's guide. Statistical Analysis System Institute Inc. Cary, NC

Schepers AR, Shanahan JF, Liebig MA, Schepers JS, Johnson SH, Luchiari A (2004) Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agron J 96:195–203

Stafford JV, Ambler B, Lark RM, Catt J (1996) Mapping and interpreting the yield variation in cereal crops. Comput Electron Agric 14:101–119

Sudduth KA, Kitchen NR, Bollero GA, Bullock DG, Wiebold WJ (2003) Comparison of electromagnetic induction and direct sensing of soil electrical conductivity. Agron J 95:472–482

Taylor JC, Wood GA, Earl R, Godwin RJ (2003) Soil factors and their influence on within-field crop variability, part II: spatial analysis and determination of management zones. Biosys Eng 84:441–453

Tominaga Y (1999) Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. Chemome Intel Lab Sys 49:105–115